315 Assignment 6

Ben Walsh

2023-05-29

```
CCC = read.csv("CCC05.csv")
ECAN = read.csv("ECAN93.csv")
```

Question 1 Generalised Additive Models

In the following Interpretations, we will not include the interpretation of the summary ie. AIC or p-values. This is because in our case an over fitted model might produce a smaller AIC and a smaller p-value, but that is not necessarily a good thing. Human interpretation is preferred over a smaller AIC or p-value, however, an excessively large AIC or p-value may be a good indicator of a bad model, this may mean it is over smoothed.

```
CCC$Date = as.Date(CCC$Date, format = "%d/%m/%Y")
ECAN$Date = as.Date(ECAN$Date, format = "%d/%m/%Y")

#Christchurch City Nitrate Levels:
CCC.gam1 = gam(Value ~ s(Date, spar=0.5), data = CCC)
CCC$predictor1 = predict(CCC.gam1)
#ggplot(CCC, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor1), #col='red')

CCC.gam2 = gam(Value ~ s(Date, spar=0.7), data = CCC)
CCC$predictor2 = predict(CCC.gam2)
#ggplot(CCC, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor2), #col='red')

CCC.gam3 = gam(Value ~ s(Date, spar=0.6), data = CCC)
CCC$predictor3 = predict(CCC.gam3)
#ggplot(CCC, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor3), #col='red')

CCC.gam4 = gam(Value ~ s(Date, spar=0.65), data = CCC)
CCC$predictor4 = predict(CCC.gam4)
#ggplot(CCC, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor4), #col='red')

CCC.gam5 = gam(Value ~ s(Date, spar=0.66), data = CCC)
CCC$predictor5 = predict(CCC.gam5)
ggplot(CCC, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor5), col='red')
```
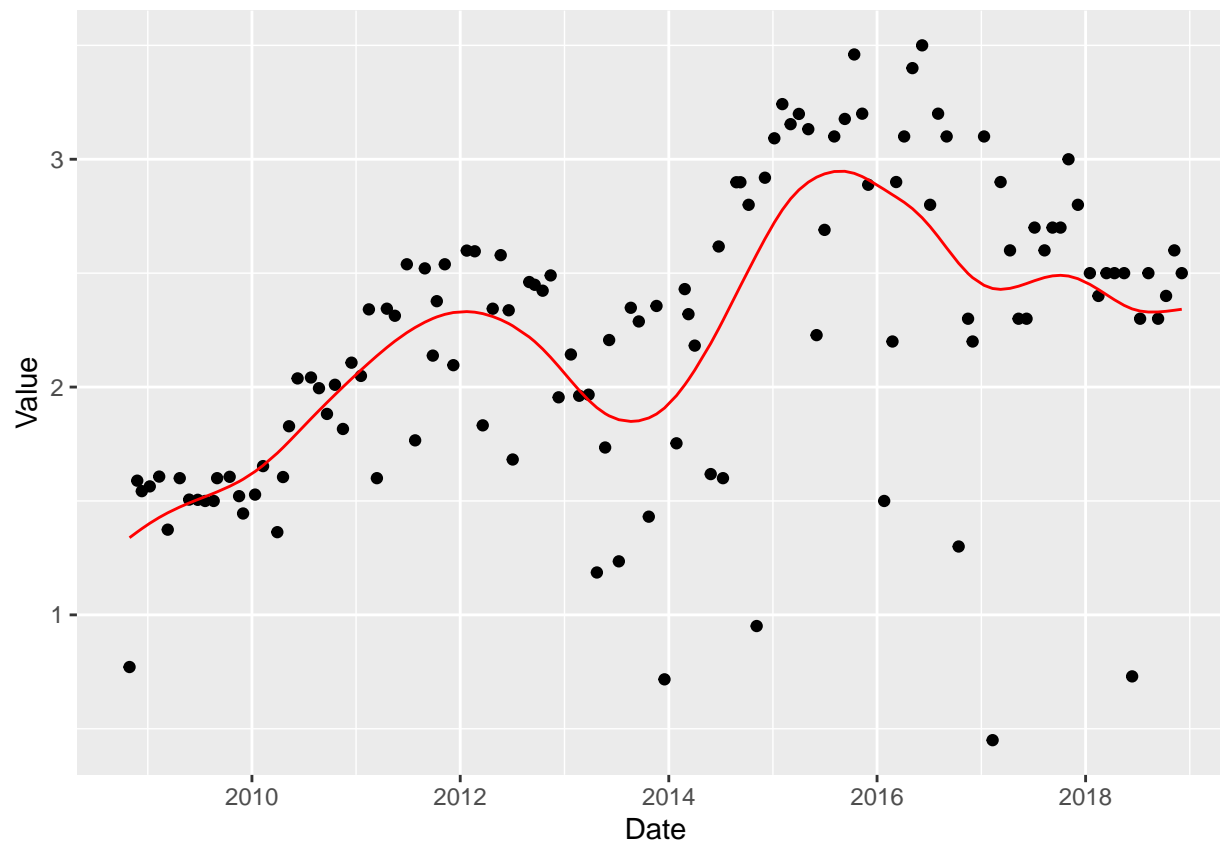
```
summary(CCC.gam5)
```

```
##
## Call: gam(formula = Value ~ s(Date, spar = 0.66), data = CCC)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98260 -0.13246  0.09806  0.26512  0.75682
##
## (Dispersion Parameter for gaussian family taken to be 0.2304)
##
##     Null Deviance: 49.5862 on 122 degrees of freedom
## Residual Deviance: 25.7835 on 111.9023 degrees of freedom
## AIC: 181.0731
##
## Number of Local Scoring Iterations: NA
##
## Anova for Parametric Effects
##                         Df Sum Sq Mean Sq F value      Pr(>F)
## s(Date, spar = 0.66)   1.0 12.219 12.2190  53.031 4.888e-11 ***
## Residuals            111.9 25.784  0.2304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                       Npar Df Npar F     Pr(F)
```

```
## (Intercept)
## s(Date, spar = 0.66)      9.1  5.526 2.564e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The data produces a sinusoidal shape over a series of years. The data has some scattered arbitriary points in the lower right quadrant of the plot, these may be due to unpredictable environmental conditions or are just random. Despite these points the data and the generalised additive model depict some sort of seasonal trend. In 2012 there is a peak in Nitrate, 2014 we experience a trough and a peak in 2015/2016 with a shallowing platue at 2017 and on-wards, overall trend is positive. The GAM model with a smoothing parameter of 0.66 is the best as it does not over smooth nor over fit, the overall trend is prominent and the model is not noisy/overfitted.

```
# Canterbury Region Nitrate Levels:
ECAN.gam2 = gam(Value ~ s(Date, spar=0.7), data = ECAN)
ECAN$predictor1 = predict(ECAN.gam2)
#ggplot(ECAN, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor1),
#col='red')


ECAN.gam3 = gam(Value ~ s(Date, spar=0.6), data = ECAN)
ECAN$predictor2 = predict(ECAN.gam3)
#ggplot(ECAN, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor2),
      #col='red')


ECAN.gam1 = gam(Value ~ s(Date, spar=0.5), data = ECAN)
ECAN$predictor3 = predict(ECAN.gam1)
#ggplot(ECAN, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor3),
#col='red')


ECAN.gam1 = gam(Value ~ s(Date, spar=0.45), data = ECAN)
ECAN$predictor4 = predict(ECAN.gam1)
ggplot(ECAN, aes(x=Date, y=Value)) + geom_point() + geom_line(aes(x=Date, y=predictor4),
col='red')
```
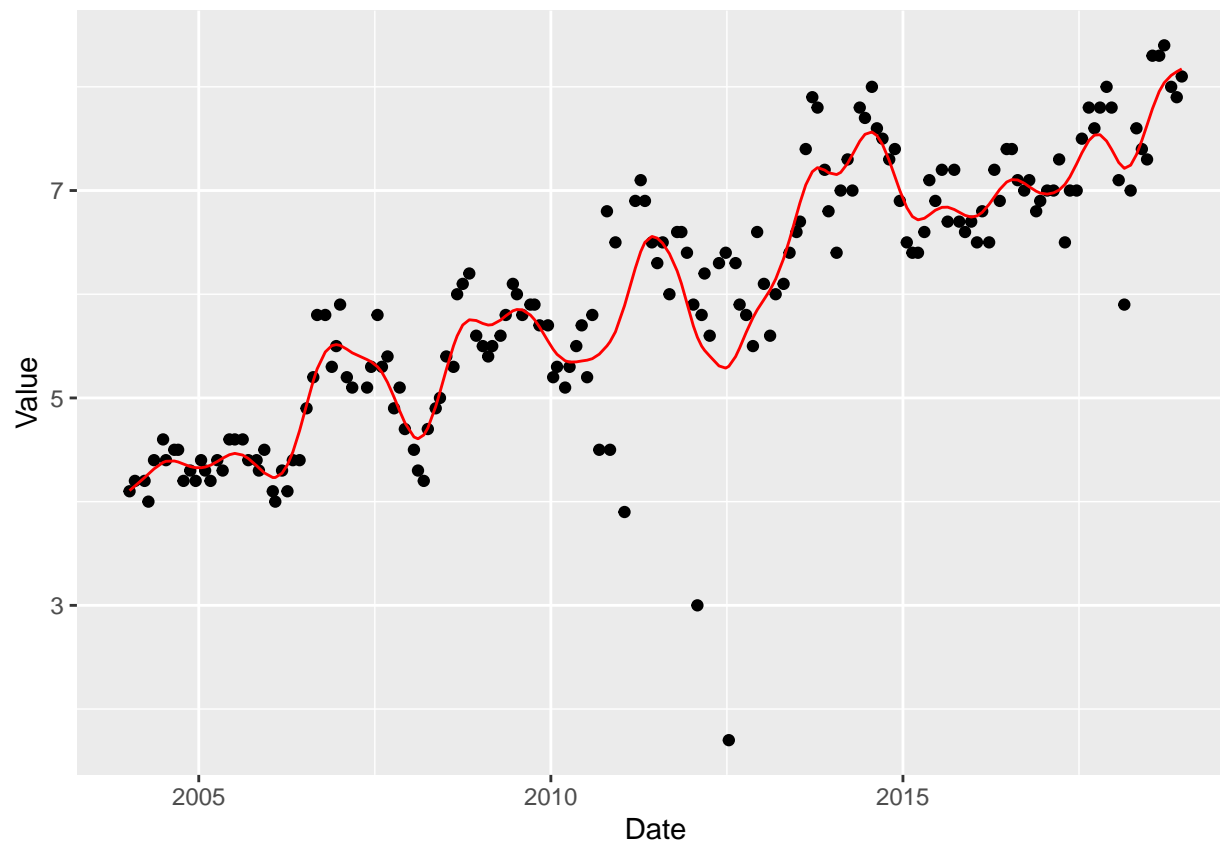
A sinusoidal shape is also apparent in our Canterbury Region. There is an odd set of points producing a trend that is almost perpendicular to our GAM trend. These points may be the result of a random environmental event/condition such as a flood, a flood would have a negative impact on the Nitrate levels in an area, thus this odd set of points and their negative trend. Despite these points the graph depicts a positive overall trend, with troughs and peaks throughout (sinusoidal). A smoothing parameter of 0.45 was chosen because there is such a quick pattern, the shape changes frequently so we need to account for the volatility by reducing the smoothing parameter. In conclusion I think that spar=0.45 is a good value as it represents the volatility accurately. After about 2015 I would consider the model over fitted as there is more of a linear trend from there on, however, if we were to increase our smoothing parameter the volatility in the majority of the graph would not be accurately represented by the GAM model, some sacrifice must be taken to represent the majority of the model.
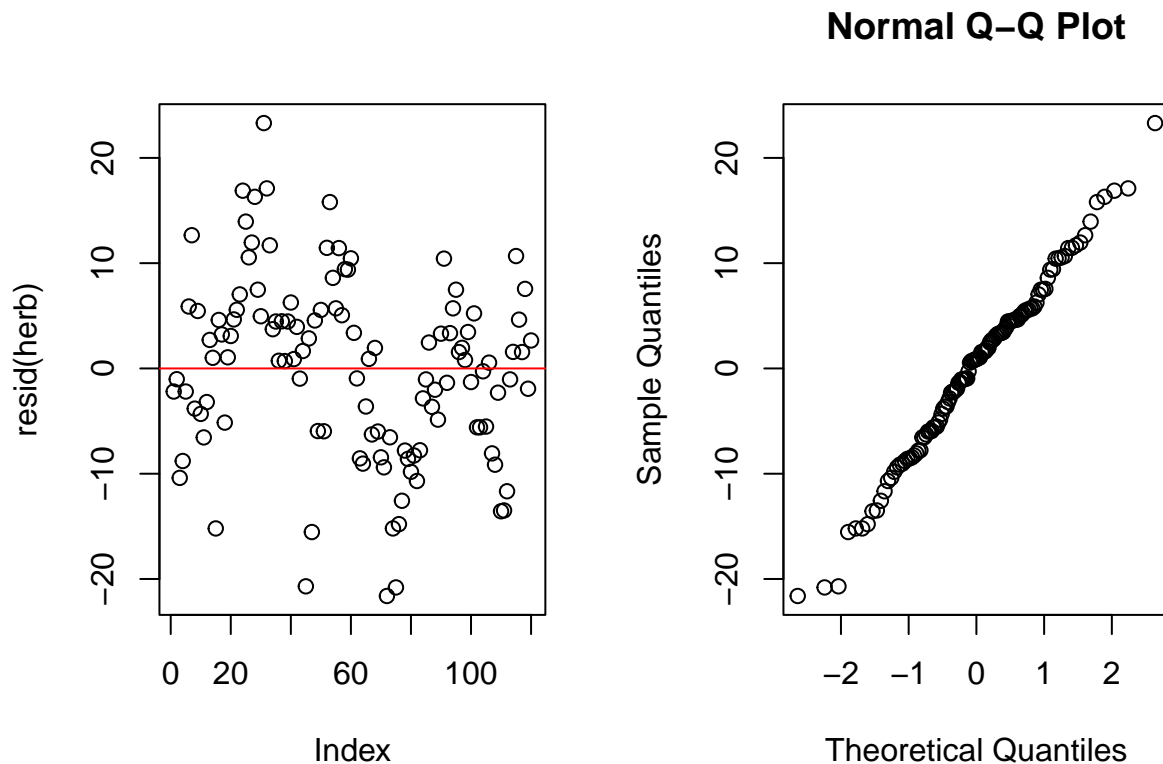
Question 2 Multiple Comparisons

Question 2a)

```
data = read_excel("Herbicides.xlsx")

herb = aov(Grass_percent ~ Herbicide, data)
summary(herb)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Herbicide    9   3092   343.5   4.412 6.09e-05 ***
## Residuals  110   8564    77.9
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(1,2))
plot(resid(herb))
abline(c(0,0), col="red")
qqnorm(herb$residuals)
```


Question 2b)

The residuals are normally distributed. Our residuals plot displays constant variance and our quantile quantile produces a linear 45º trend. This depicts homoscedasticity and normality, main assumptions we make about our residuals.

Question 2c)

```
#LSD test
mse <-sum(herb$residuals*herb$residuals)/herb$df.residual
LSD.test(data$Grass_percent, data$Herbicide, herb$df.residual, mse, console=TRUE)
```

```
##
## Study: data$Grass_percent ~ data$Herbicide
```

```
## 
## LSD t Test for data$Grass_percent
## 
## Mean Square Error:  77.8534
## 
## data$Herbicide,  means and individual ( 95 %) CI
## 
##                          data.Grass_percent       std  r      LCL      UCL     Min
## Aminopyralid                        63.44375  9.913055 12 58.39597 68.49153 49.875
## Aminopyralid+triclopyr              62.89583  8.645617 12 57.84805 67.94361 52.500
## Chlorsulfuron                       52.77083  5.158244 12 47.72305 57.81861 44.500
## Flumetsulam                         58.09375  6.201202 12 53.04597 63.14153 49.500
## MCPA                                58.31875  8.093657 12 53.27097 63.36653 45.750
## MCPB                                55.11458 10.260590 12 50.06680 60.16236 33.500
## MCPB+bentazone                      52.29167  8.893201 12 47.24389 57.33945 36.750
## Nil                                 52.04167  7.303551 12 46.99389 57.08945 40.375
## Sclerotinia                         47.19792 12.355696 12 42.15014 52.24570 26.500
## Thifensulfuron-methyl               50.30208  9.196476 12 45.25430 55.34986 29.500
##                            Max
## Aminopyralid              86.75
## Aminopyralid+triclopyr    80.00
## Chlorsulfuron             60.25
## Flumetsulam               70.75
## MCPA                      69.75
## MCPB                      72.00
## MCPB+bentazone            64.25
## Nil                       63.50
## Sclerotinia               63.50
## Thifensulfuron-methyl     64.25
## 
## Alpha: 0.05 ; DF Error: 110
## Critical Value of t: 1.981765
## 
## least Significant Difference: 7.138638
## 
## Treatments with the same letter are not significantly different.
## 
##                          data$Grass_percent groups
## Aminopyralid                        63.44375      a
## Aminopyralid+triclopyr              62.89583      a
## MCPA                                58.31875     ab
## Flumetsulam                         58.09375     ab
## MCPB                                55.11458     bc
## Chlorsulfuron                       52.77083    bcd
## MCPB+bentazone                      52.29167    bcd
## Nil                                 52.04167    bcd
## Thifensulfuron-methyl               50.30208     cd
## Sclerotinia                         47.19792      d
```

The groups with the same letters under the column "groups" are the herbicides that are not significantly different from one another. If a herbicide does not include a letter from another herbicide then those two are significantly different as their Grass_percent is different by at least 7.1386, our least significant difference constant. There is 17 groups with significant difference, it seems over the top to list them all.

Question 2d)

```
#Bonferroni adjustment
pairwise.t.test(data$Grass_percent, data$Herbicide, p.adj = "bonferroni",console=TRUE)
```

```
##
##   Pairwise comparisons using t tests with pooled SD
##
## data:   data$Grass_percent and data$Herbicide
##
##                        Aminopyralid Aminopyralid+triclopyr Chlorsulfuron
## Aminopyralid+triclopyr 1.00000      -                      -
## Chlorsulfuron          0.16810      0.26332                -
## Flumetsulam            1.00000      1.00000                1.00000
## MCPA                   1.00000      1.00000                1.00000
## MCPB                   1.00000      1.00000                1.00000
## MCPB+bentazone         0.11201      0.17800                1.00000
## Nil                    0.09018      0.14438                1.00000
## Sclerotinia            0.00073      0.00133                1.00000
## Thifensulfuron-methyl  0.01824      0.03068                1.00000
##                        Flumetsulam MCPA    MCPB    MCPB+bentazone Nil
## Aminopyralid+triclopyr -           -       -       -              -
## Chlorsulfuron          -           -       -       -              -
## Flumetsulam            -           -       -       -              -
## MCPA                   1.00000     -       -       -              -
## MCPB                   1.00000     1.00000 -       -              -
## MCPB+bentazone         1.00000     1.00000 1.00000 -              -
## Nil                    1.00000     1.00000 1.00000 1.00000        -
## Sclerotinia            0.13938     0.11505 1.00000 1.00000        1.00000
## Thifensulfuron-methyl  1.00000     1.00000 1.00000 1.00000        1.00000
##                        Sclerotinia
## Aminopyralid+triclopyr -
## Chlorsulfuron          -
## Flumetsulam            -
## MCPA                   -
## MCPB                   -
## MCPB+bentazone         -
## Nil                    -
## Sclerotinia            -
## Thifensulfuron-methyl  1.00000
##
## P value adjustment method: bonferroni
```

From the Bonferroni adjustment there are 4 significantly different pairs of Herbicide. Sclerotinia is significantly different to Aminopyralid and Aminopyralid+triclopyr as they produce a p-value of 0.00073 and 0.00133 respectively, less than a significance level of 0.05. Thifensulfuron-methyl is significantly different to both Aminopyralid and Aminopyralid+triclopyr as they produce a p-value of 0.01824 and 0.03068 respectively, less than a significance level of 0.05.

```
#Tukey HSD
dataHSD <- TukeyHSD(aov(Grass_percent ~ Herbicide, data))
plot(dataHSD)
```

95% family-wise confidence level

Differences in mean levels of Herbicide
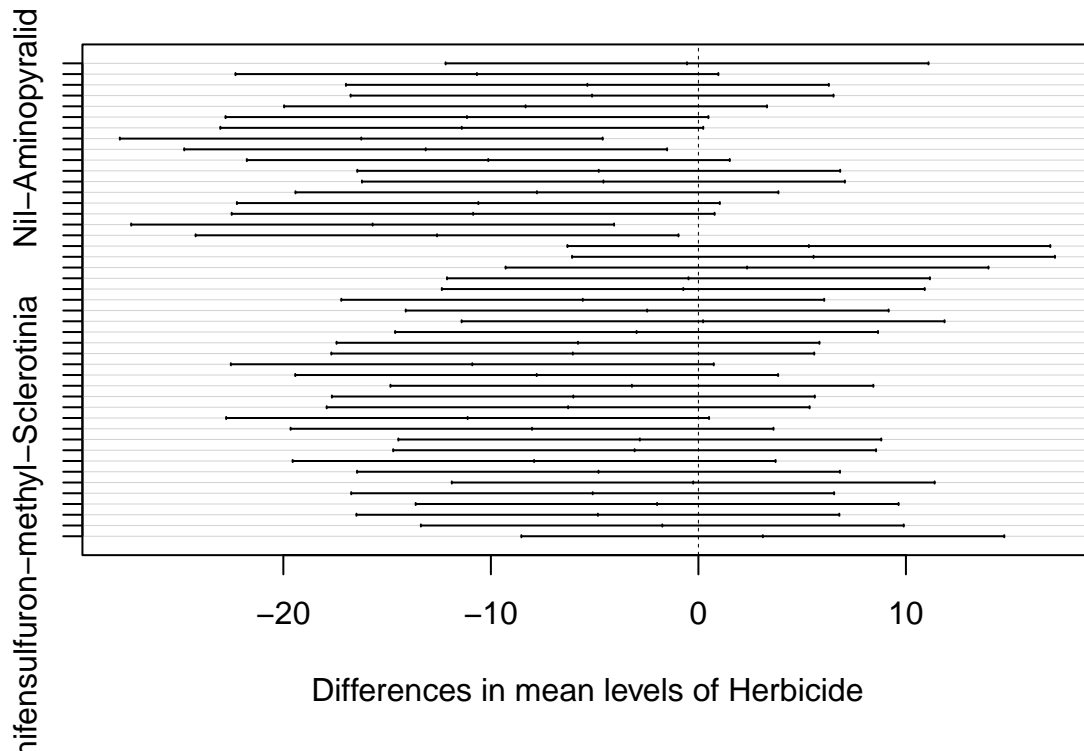
```
Tukeyaov <- aov(Grass_percent ~ Herbicide, data)
summary(Tukeyaov)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Herbicide      9   3092   343.5   4.412 6.09e-05 ***
## Residuals    110   8564    77.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
dataHSD <- TukeyHSD(Tukeyaov)
```

The following pairs of herbicide produce p-values < 0.05, indicating that there is a significant difference between the two.

- Sclerotinia-Aminopyralid+triclopyr , p-value = 0.0011879

- Thifensulfuron-methyl-Aminopyralid+triclopyr , p-value = 0.0228998

- Sclerotinia-Aminopyralid , p-value = 0.0006647

- Thifensulfuron-methyl-Aminopyralid , p-value = 0.0142167

All 4 of these are the same significant pairs as the bonferroni adjustment.

```
## Dunnett test
data$Trta <- as.factor(data$Herbicide)
herb2 <- aov(Grass_percent ~ Trta, data)
test.dunnett=glht(herb2, linfct=mcp(Trta="Dunnett"))
confint(test.dunnett)
```

```
##
##      Simultaneous Confidence Intervals
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: aov(formula = Grass_percent ~ Trta, data = data)
##
## Quantile = 2.7308
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                                              Estimate  lwr       upr
## Aminopyralid+triclopyr - Aminopyralid == 0   -0.5479 -10.3846    9.2888
## Chlorsulfuron - Aminopyralid == 0           -10.6729 -20.5096   -0.8362
## Flumetsulam - Aminopyralid == 0              -5.3500 -15.1867    4.4867
## MCPA - Aminopyralid == 0                     -5.1250 -14.9617    4.7117
## MCPB - Aminopyralid == 0                     -8.3292 -18.1659    1.5075
## MCPB+bentazone - Aminopyralid == 0          -11.1521 -20.9888   -1.3154
## Nil - Aminopyralid == 0                     -11.4021 -21.2388   -1.5654
## Sclerotinia - Aminopyralid == 0             -16.2458 -26.0825   -6.4091
## Thifensulfuron-methyl - Aminopyralid == 0   -13.1417 -22.9784   -3.3050
```

```
op <- par()
par(mar=c(5,10,4,2))
plot(test.dunnett)
```

95% family−wise confidence level
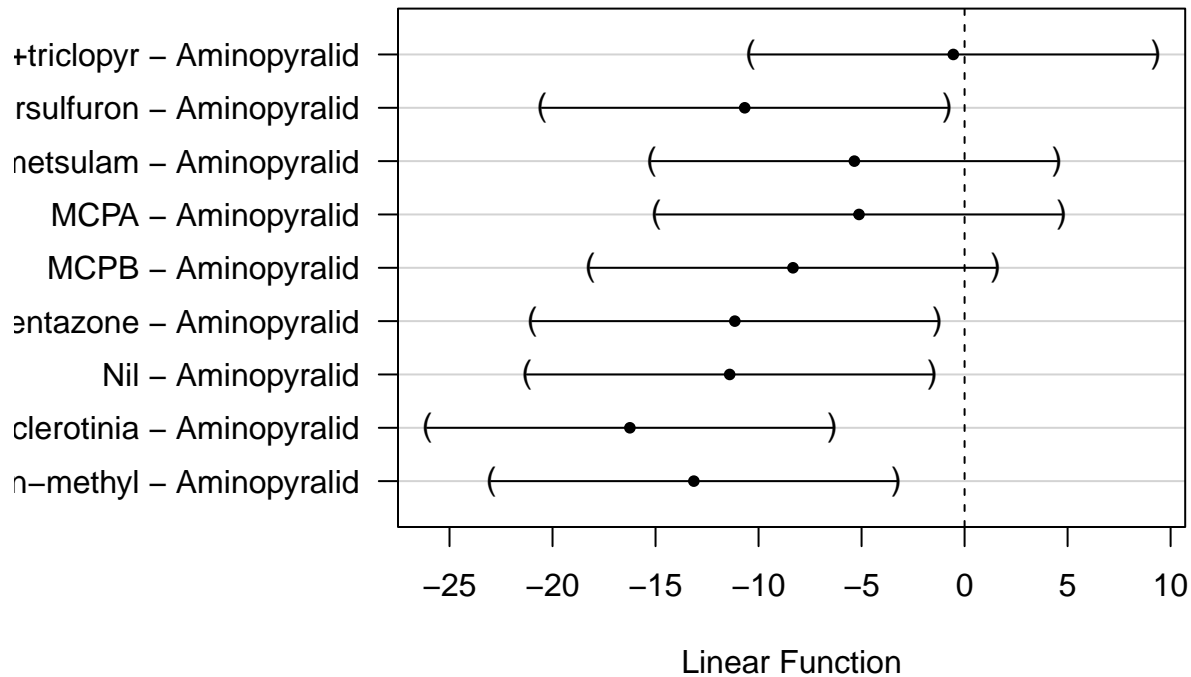
Using our plot of our Dunnett test, it is apparent that 5 herbicides paired with our response herbicide "Aminopyralid" are of significance. This is because they do not fall on 0, (do not have 0 within their confidence intervals), which suggests that there is significant difference between those herbicides. Those herbicides are:

- Chlorsulfuron - Aminopyralid
- MCPB+bentazone - Aminopyralid
- Nil - Aminopyralid
- Sclerotinia - Aminopyralid
- Thifensulfuron-methyl - Aminopyralid

Question 2e)

The conclusion in c was that there were 17 paired herbicides with significant differences. However, the three conclusions from d) only differentiate 4, 4, 5 significant herbicide pairs from the entire set respectively. As the statistician I would trust the tests made in d over the LSD test, and I can confidently conclude that the herbicides:

- Sclerotinia-Aminopyralid
- Thifensulfuron-methyl-Aminopyralid

are significantly different as in all three tests made in d), these above two pairs were denoted as significantly different.The Herbicides:

- Sclerotinia-Aminopyralid+triclopyr

- Thifensulfuron-methyl-Aminopyralid+triclopyr

I would consider are also significantly different as they both appear in the TukeyHSD and Bonferroni tests, where the Dunnett test only includes those herbicides paired with Aminopyralid. At the end of the day it is down to what makes the most sense in context, some information from the farmers would help us decide the most ideal herbicide and it would also help us decide which test is best for such situation.